Biomedical Semantic Data Resources for Data Analytics and Knowledge Discovery

Abstract

Semantic resources play an increasingly important role in the knowledge discovery in the biomedical domain. The biomedical community (see Elixir network) is developing towards an integrated infrastructure where all data sources are interoperable, making use of Semantic Web technologies.

This tutorial is a hands-on experience for scientists who use the biomedical data and want to learn how to efficiently exploit the available semantic data resources (UniProtKb, PubChem, BioOpener, LinkedPPI and others). This tutorial makes reference to the data resources under development from the Elixir community which are developed in order to satisfy the domain experts requirements.

In the biomedical domain the integration and transfer of knowledge between textual resources (i.e. the scientific literature) and knowledge bases (e.g., Uniprot KB) is an ongoing process which is tackling challenges such as data standardization, efficient data representation and data interoperability. Semantic Web solutions have been adopted in recent years (i.e., Semantic Web and Linked Data technologies) to overcome these hurdles. Despite the growing interest in semantics driven data resources, the biomedical scientific community does not fully exploit the advantages of the semantics driven data integration to achieve better data sharing, information extraction and knowledge discovery.

This tutorial covers all aspects of semantics solutions in the biomedical domain and the benefits from existing data sources. Specific focus is set to data standardization, exploitation of data resources for scientific approaches and the prospects of Semantic Web technologies.

Audience Profile

This tutorial is designed in a way that targets wide range of researchers based on their expertise and interests including:

- Computer scientists who are interested in using biomedical semantic resources.
- Bioinformaticians with specific questions for exploiting semantic knowledge bases.
- Biologists who are interested in using available tools and frameworks.

The tutorial provides the use of semantic resources in theory and practice. It accounts for hands-on use of existing resources, teaches access and reuse of data and will provide an outlook into resource-driven knowledge discovery.

Tutorial outline

1. Introduction to Semantic Web (1 hour and a half)

In this session we describe the fundamental concepts of Semantic Web for the audience with less or no knowledge of Semantic Web.

- Introduction to Semantic Web
- Introduction to RDF and triple stores
- Data standardization approaches
- Data interoperability and information exchange
- Use of ontologies for data interoperability
- Retrieval of data based on terms, concepts and URIs

2. SPARQL Federation Approaches (1 hour and a half)

In this session we discuss the concepts around SPARQL query Federation to access multiple heterogeneous biological data sets to draw meaningful biological correlations. Real life sciences dataset (e.g Drugbank, Dailymed) will be queried to elaborate SPARQL Federation. We cover the following topics:

- Introduction to the SPARQL Endpoint Federation (SEF)
- Linked Data Federation (LDF)
- Distributed Hash Tables (DHTs)
- Hybrids of SEF+ LDF
- SPARQL Query Federation Engines
- Formulating SPARQL queries on distributed data resources, advantages and limitations, comparison against relational databases
- Use of ontological concepts (and biomedical terminologies) in query federation

3. Knowledge discovery based on Semantic Web solutions (1 hour and a half)

In this session we focus on retrieving relevant scientific information and identifying connections between pieces of scientific knowledge. We focus on automatic literature analysis and its integration with biomedical data resources. We touch upon the latest developments in bio text mining which make use of Semantic Web resources.

- Knowledge discovery (KD) based on Semantic Web solutions
- Use of semantic resources in data analytics and workflow systems
- Approaches of knowledge discovery, i.e. co-evaluation of public data resources
- Combining Semantic Web solutions with literature resources
- Known solutions for multi-repository data retrieval and its KD benefits

4. In silico to in vivo with Linked Data (1 hour and a half)

This session gives an overview of available tools which have been developed for domain experts (e.g. biologist) in which Linked Data concepts were used in order to integrate multiple heterogeneous biomedical data resources.

- Introduction to available resources and tools.
- Overview of LinkedPPI
- Overview of BioOpener

Tutorial Speakers

Prof Dietrich Rebholz-Schuhmann is established chair for data analytics at the National University of Ireland, Galway, and the director of the Insight Centre for Data Analytics in Galway. His research is positioned in semantic technologies in the biomedical domain including biomedical informatics, literature analysis, ontologies, Semantic Web and Information representation. He is also editor-in-chief of the Journal of Biomedical Semantics.

Dr. Ratnesh Sahay is a research fellow and leading the Semantics in eHealth and Life Sciences (SeLS) research unit at the Insight Centre for Data Analytics, National University of Ireland, Galway. He is the lead scientist of BioOpener project and his research emphasis on using semantics for solving key integration/interoperability challenges in the e-health, clinical trial and biomedical domains. He is a

member of the Global Alliance for Genomics and Health, Health Level Seven (HL7) Standard and World Wide Web Consortium (W3C) standardization working groups (OWL, HCLS). He previously served as a member of the OASIS SEE Technical Committee, W3C SWS-Challenge working group and CMS Working Group.

Dr. Ali Hasnain is a postdoctoral researcher at Insight Centre for Data Analytics in Galway. His research interests include Linked Open Data Big Data, Semantic Models, Data Cataloguing and Linking, Semantic Matching and Relatedness, Link Discovery, Visual Query Formulation, Data Provenance and Data Integration. He is the program committee member of different international workshops e.g. VOILA at ISWC and conferences e.g. KESW. Moreover he has been involved in organising different workshops and tutorial e.g. at K-Cap 2015 and SWAT4LS 2015 for international audiences.

Dr. Laleh Kazemzadeh is a postdoctoral research at Royal College of Surgeons in Ireland, Dublin and associated member of Insight Centre for Data Analytics in Galway. Her research interest focuses on application of Semantic Web and Linked Data in integration of heterogeneous biomedical data. She is the lead scientist on LinkedPPI project.